Application Of Missing Feature Theory To The Recognition Of Musical Instruments In Polyphonic Audio

Jana Eggink and **Guy J. Brown**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
{j.eggink, g.brown}@dcs.shef.ac.uk

Abstract

A system for musical instrument recognition based on a Gaussian Mixture Model (GMM) classifier is introduced. To enable instrument recognition when more than one sound is present at the same time, ideas from missing feature theory are incorporated. Specifically, frequency regions that are dominated by energy from an interfering tone are marked as unreliable and excluded from the classification process. The approach has been evaluated on clean and noisy monophonic recordings, and on combinations of two instrument sounds. These included random chords made from two isolated notes and combinations of two realistic phrases taken from commercially available compact discs. Classification results were generally good, not only when the decision between reliable and unreliable features was based on the knowledge of the clean signal, but also when it was solely based on the harmonic overtone series of the interfering sound.

1 Introduction

Music transcription describes the process of finding a symbolic representation for a piece of music based on an audio recording or possibly a live performance. A symbolic representation in this context generally means some kind of musical score, with information for every tone about its fundamental frequency (F0), its onset time and duration, the instrument on which the tone was played, and possibly loudness and other expressive gestures. Transcription is a task that is currently almost exclusively performed by trained musicians; computer based automatic transcription remains a challenging problem. In the present study we focus on one part of the automatic music transcription problem - instrument recognition from an audio recording.

Realistic sound recordings from commercially available compact discs (CDs) have been successfully used in systems limited to monophonic sound recognition. Martin (1999) used a number of features related to both temporal and spectral characteristics of instrument sounds in a hierarchical classification scheme. Generally, the performance of his system was comparable to human performance, although humans outperformed the computer system in instrument family differentiation. Using 27 different instruments, the system achieved a recognition accuracy of 57% for realistic monophonic examples and 39% for isolated tones with the best possible parameter settings. Reducing the number of instruments to 6 improved results up to 82% for monophonic phrases.

Brown *et al.* (2001) described a classifier based on Gaussian mixture models (GMMs), and compared the influence of different features on classification accuracy. Test material consisted of realistic monophonic phrases from four different woodwinds. Both cepstral features and features related to spectral smoothness performed well. With these features they achieved an average recognition accuracy of around 60%, reaching 80% for the best possible parameter combination and choice of training material.

Marques and Moreno (1999) compared the performance of classifiers based on Gaussian mixture models and support vector machines (SVMs). Cepstral features performed better than linear prediction based features; and mel-frequency scaled cepstral features performed again better than linearly scaled ones. Using realistic recordings of 8 different instruments, they achieved recognition accuracies of 63% for the GMM-based classifier and a slightly improved result of 70% for SVMs, but the influence of the choice of features seemed to be higher than that of the classification method.

Only very few studies have attempted instrument recognition for polyphonic music, and the systems were mostly tested on very limited and artificial examples. Kashino and Murase (1999) used a template-based time domain approach. For each note of each possible instrument an example waveform was stored. As a first step, the sound file was divided according to onsets. For every part the most prominent instrument tone was then determined by comparing the mixture with the phase-adjusted example waveforms. In an iterative processing cycle, the energy of the corresponding waveform was subtracted to find the next most prominent instrument tone. Using only three different instruments (flute, violin and piano) and specially arranged ensemble recordings they achieved 68% correct instrument identifications with both the true F0s and the onsets supplied to the algorithm. With the inclusion of higher level

musical knowledge, most importantly voice leading rules, recognition accuracy improved to 88%.

A frequency domain approach was proposed by Kinoshita *et al.* (1999), using features related to the sharpness of onsets and the spectral distribution of partials. F0s were extracted prior to the instrument classification process to determine where partials from more than one F0 would coincide. Corresponding feature values were either completely ignored or used only after an average value corresponding to the first identified instrument was subtracted. Using random two-tone combinations from three different instruments (clarinet, violin, piano), they obtained recognition accuracies between 66% and 75% (73%-81% if the correct F0s were provided), depending on the interval between the two notes.

In this paper, we propose an approach based on missing feature (or missing data) theory to enable instrument recognition in situations where multiple tones may overlap in time. The general idea is to use only the parts of the signal which are dominated by the target sound, and ignore features that are dominated by background noise or interfering tones. This approach is motivated by a model of auditory perception which postulates a similar process in listeners; since target sounds are often partially masked by an interfering sound, it can be inferred that listeners are able to recognize sound sources from an incomplete acoustic representation (Cooke *et al.*, 2001). The missing feature approach has previously been successfully applied in the fields of robust speech recognition (Cooke *et al.*, 2001) and speaker identification (Drygajlo and El-Maliki, 1998), the latter task beeing one which is closely related to musical instrument identification.

In polyphonic music, partials of one tone often overlap with those of another tone. As a consequence, the energy values of these partials no longer correspond to those of either instrument, and most existing instrument recognition techniques will fail. Within a missing feature approach, these corrupted features will be excluded from the recognition process. The remaining information will therefore be incomplete, but feature values will mainly contain information about one sound source only. The hope is that this remaining information is still sufficient to enable robust instrument classification.

The main requirement for the actual classifier is its robustness towards incomplete feature sets. Classifiers based on Gaussian mixture models (GMMs) can be easily adapted to work with incomplete data (Drygajlo and El-Maliki, 1998). They have also been successfully employed for instrument classification in monophonic music (Brown *et al.*, 2001; Marques and Moreno, 1999) and are therefore a promising choice for a system attempting instrument classification for polyphonic music.

2 System Description

A schematic view of our system is shown in Figure 1. The first stage is a frequency analysis of the sampled audio signal. Subsequently, the F0s of all tones are extracted and frequency regions where partials of a non-target tone are found are marked as unreliable. Hence, a binary 'mask' is derived, that indicates the features which should be employed by a GMM classifier.

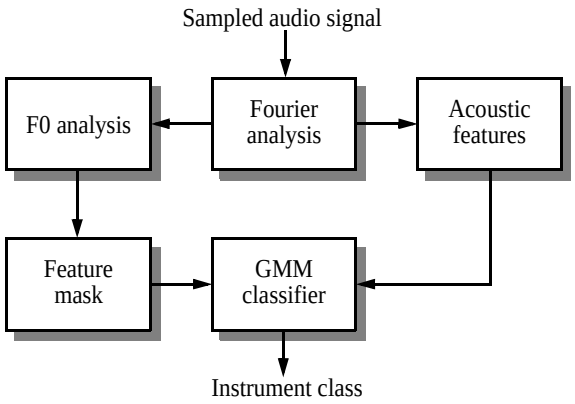

Figure 1: Schematic of the instrument classification system.

2.1 Acoustic Features

The choice of acoustic features is very important for any classification system. While cepstral features, especially when mel-frequency scaled, have been proven to give good results for musical instrument classification systems (see section 1), they do not easily fit within a missing feature approach. The idea of the missing feature approach is to exclude frequency regions dominated by energy from an interfering sound source. A specific frequency region does not have a clear correspondence in the cepstral domain, so that a distinction between features dominated by the target tone and those dominated by an interfering tone cannot be made. Therefore local spectral features are required for the missing feature approach.

From these considerations, we chose linearly scaled features over a quasi-logarithmic scaling which would be closer to human hearing. The harmonic overtone series of musical tones is approximately evenly spaced on a linear scale, and an equally linear scaling of features makes it easier to block out the energy of such an interfering harmonic series.

The employed features can basically be described as a coarse spectrogram. Sampled audio recordings were divided into frames 40 ms in length with a 20 ms overlap. Each frame was multiplied with a Hanning window, and a fast Fourier transform (FFT) was computed. The resulting spectra were log compressed and normalised to a standard maximum value. Each feature consists of the spectral energy within a 60 Hz wide frequency band. The features span a region between 50 Hz and 6 kHz, with 10 Hz overlap between adjacent features, resulting in a total of 120 features per time-frame. The overall frequency range includes all possible F0s of the instruments used, and their formant regions.

2.2 Fundamental Frequency Detection

The system described here depends on the estimation of F0s prior to any instrument identification. To evaluate the accuracy of the instrument classification independent from a pitch detection system, we decided to circumvent the problem by

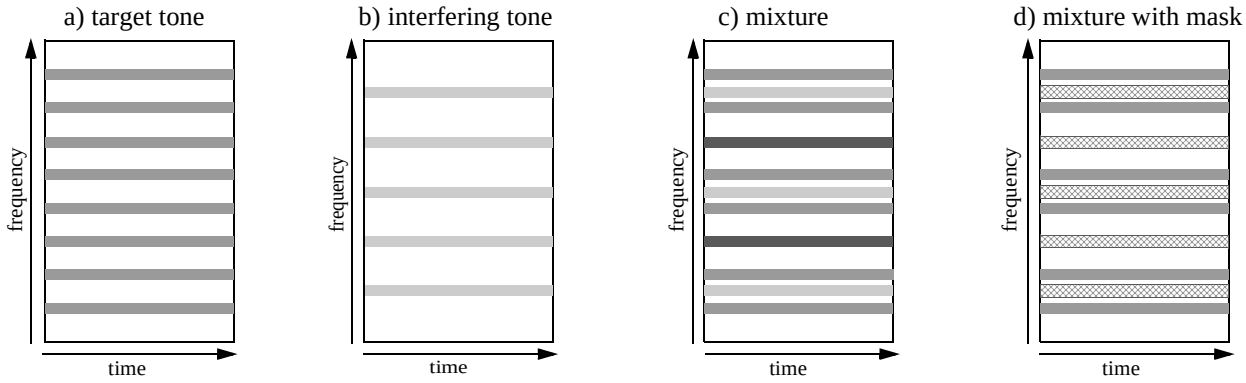

Figure 2: Example for missing feature masks. Simplified spectra of a) the target tone, b) the interfering tone, and c) the mixture of both tones. Energy values which, due to overlapping partials, do not correspond to those of either tone alone are shown in dark grey. In d) the mixture is overlaid with the mask, represented by hatched bars.

using either *a priori* masks (see section 2.3.1) or isolated tones with a known F0 that could be manually supplied to the system.

Finding multiple F0s in polyphonic music is known to be a non-trivial problem, and a growing number of publications is focusing on its various aspects (e.g. see Klapuri, 2001; Raphael, 2002), with encouraging results if the number of concurrent voices is low. In an earlier publication (Eggink and Brown, 2003), we presented an iterative pattern matching approach based on 'harmonic sieves'. While no extensive tests were carried out, the results were generally good for two-voiced music. The advantage of this approach lies in its explicit identification of spectral peaks which belong to the harmonic overtone series of the different F0s. If the real frequency location of partials is known, the assumption of an exactly harmonic overtone spcectrum can be dropped and more accurate missing feature masks can be derived.

2.3 Feature Masks

2.3.1 A priori Masks

Feature masks are used to indicate which features should be used for the classification process. The identification of these reliable features is often one of the hardest problems in missing feature systems. Commonly, *a priori* masks are used to establish an upper performance limit. If the clean signal (i.e. the monophonic target signal alone without any interfering noise or any other sound sources) is known, it can be compared with the mixture, and only those parts where the mixture is similar to the target sound alone are used for recognition. Feature values were computed for the target sound alone and for the mixture consisting of the target and the interfering sound. They were marked as reliable only when then the feature value of the mixture was within a range of ±3dB compared to the corresponding feature value of the clean signal. The threshold of ±3 dB is somewhat arbitrary, but led to good results in initial studies, and a lower threshold of ±1 dB gave generally similar results.

2.3.2 Pitch-based Masks

While *a priori* masks provide a good tool to assess a best possible performance, they are not very realistic, as the clean signal is not normally available. A more realistic way to generate the missing feature masks is based on the F0 of the interfering tone (or possibly tones). The energy of harmonic tones is concentrated in their partials, whose positions can be approximated once the F0 is known. If a partial from the non-target tone falls within the frequency range of a feature, the feature is marked as unreliable and not used for recognition. This approach obviously depends on the harmonic structure of the interfering tone, and is therefore suitable for most musical instrument tones, but does not work for percussion and other inharmonic sounds. In such cases, other cues (such as e.g. stereo position) could be used.

2.4 Gaussian Mixture Model Classifier with Missing Features

A GMM models the probability density function (pdf) of observed features by a multivariate Gaussian mixture density:

$$p(x) = \sum_{i=1}^{N} p_i \Phi_i(x, \mu_i, \Sigma_i) \qquad (1)$$

where x is a D-dimensional feature vector and N is the number of Gaussian densities Φ_i, each of which has a mean vector μ_i, covariance matrix Σ_i and mixing coefficient p_i. Here, we assume a diagonal covariance matrix; although this embodies an assumption which is incorrect (independence of features) it is a widely used simplification (e.g., see Brown *et al.*, 2001). Accordingly, (1) can be rewritten as:

$$p(x) = \sum_{i=1}^{N} p_i \prod_{j=1}^{D} \Phi_i(x_j, m_{ij}, \sigma_{ij}^2) \qquad (2)$$

where m_{ij} and σ_{ij}^2 represent the mean and variance respectively of a univariate Gaussian pdf. Now, consider the case in which some components of x are missing or unreliable, as indicated by a binary mask M. In this case, it can be shown (Drygajlo and El-Maliki, 1998) that the pdf (2) can be computed from partial data only, and takes the form:

$$p(x_r) = \sum_{i=1}^{N} p_i \prod_{j \in M} \Phi_i(x_j, m_{ij}, \sigma_{ij}^2) \qquad (3)$$

where *M'* is the subset of reliable features x_r in *M*. Hence, missing features are effectively eliminated from the computation of the pdf.

2.5 Bounded Marginalisation

With the binary masks described so far, all information from the features marked as unreliable is completely discarded. But the features still hold some information, as the observed energy value represents an upper boundary for the possible value of the target sound (Cooke *et al.*, 2001). Instead of ignoring the unreliable features, the pdf can be approximated as a product of the activation based on the reliable features x_r and an integration over all possible values of the unreliable features x_u:

$$p(x_r, x_u) = \sum_{i=1}^{N} p_i \Phi_i(x_r, \mu_i, \Sigma_i) \int \Phi_i(x_u, \mu_i, \Sigma_i) dx_u \qquad (4)$$

If the upper and lower bounds (x_{high}, x_{low}) of the unreliable features are known, for diagonal covariance matrices the integral can be evaluated as a vector difference of multivariate error functions (Cooke *et al.*, 2001). Since no specific knowledge exists for the lower boundary, it is always assumed to be zero and the corresponding error function is subsequently ignored. The integral in (4) can then be computed as:

$$\int \Phi_i(x_u, \mu_i, \Sigma_i) dx_u = \frac{1}{2}\left[erf\left(\frac{x_{high,\,u} - \mu_{u,\,i}}{\sqrt{2\sigma^2_{u,\,i}}} \right) \right] \qquad (5)$$

where $x_{high,u}$ represents the upper bound of the unreliable feature x_u, and $\mu_{u,i}$ and $\sigma^2_{u,i}$ the mean and variance respectively of the unreliable feature of centre *i*.

2.6 Training

Individual GMMs were trained for five different instruments (flute, oboe, clarinet, violin and cello). To make the models as robust as possible they were trained with different recordings for each instrument, using both monophonic musical phrases and single tone recordings. After an initial clustering using a K-means algorithm, the parameters of the GMMs were trained by the expectation-maximisation (EM) algorithm. The number of Gaussian densities, *N*, was set to 120 after some experimentation; a further increase gave no improvement.

3 Evaluation

Both realistic phrases from commercially available CDs and isolated samples were used for evaluation purposes. The advantage of the former is that it is closer to realistic applications, and likely to include a range of acoustic properties that can pose additional difficulties to a recognition system, like e.g. reverberation and a wide range of tempo and dynamic differences. Isolated samples on the other hand make it possible to evaluate a system independent of a pitch extractor, as the F0s are known beforehand. It also allows systematic testing of specific chord combinations.

3.1 Monophonic Sounds

To establish an upper limit on performance with missing features, tests were carried out with monophonic recordings. Test material was taken from recordings which were not included in the training material, consisting of chromatic scales from the McGill master samples CD (Opolko and Wapnick, 1987), the Ircam studio online collection, and the Iowa musical instrument samples. Different models were trained by a leave-one-out cross validation scheme, each using only two of the mentioned sample collections, but the same realistic monophonic phrases from commercially available classical music CDs. To avoid cues based solely on the different pitch range of the instruments, only tones from one octave (C4-C5) were used for testing, although the models were always trained on the full pitch range of the instruments. Where necessary, chromatic scales were manually cut into single tones.

Classification decisions were made for each frame independently and the model which accumulated the most 'wins' over the tone or phrase duration was taken as the overall classification for that example. Average instrument recognition accuracy was 66% for the McGill samples, 70% for the Ircam samples and 62% for the Iowa samples (the last one excluding the violin, for which no recordings were available). A confusion matrix averaing across all three conditions is shown in Table 1.

response / stimulus	Flute	Clarinet	Oboe	Violin	Cello
Flute	**67%**	8%	0%	15%	8%
Clarinet	23%	**59%**	5%	8%	5%
Oboe	0%	10%	**85%**	3%	3%
Violin	4%	4%	8%	**65%**	19%
Cello	3%	10%	15%	15%	**56%**

Table 1: Confusion matrix for mean instrument recognition of single notes.

Identification performance was also assessed on monophonic phrases from a number of classical music CDs, which were not used for training. For every instrument 5 different recordings of varying length from 2-10 seconds were used, with classification decisions made for each sound file separately. Results were very similar for the 3 sets of models, with an average recognition accuracy of 88%. All flute, clarinet and violin examples were correctly classified, while up to 2 oboe examples were mistaken for flutes and up to 2 cello examples were confused with clarinets.

The main reason for the better classification of realistic phrases seems to lie in their generally longer duration and higher variability. If one tone of a certain F0 gets misclassified, this is more likely to be evened out by a majority of correctly classified tones. If the results are compared on a frame by frame basis, the system performs equally well on realistic phrases and single notes, with an average of 60% correctly classified frames.

3.2 *A Priori* Masks

A priori masks provide a good tool to accurately select reliable and unreliable features if the clean signal is available. The performance achieved can be regarded as an upper limit, and a starting point for more realistic methods of distinguishing between reliable and unreliable features.

3.2.1 Noise

The system was shown earlier to be very robust against random deletions of features (Eggink and Brown, 2003). Here we test its robustness towards artificially added noise with and without missing feature masks. Aside from providing a good evaluation method for the missing features approach, noise robustness may be relevant in cases where low quality recordings are transcribed, such as live performances or old analog records.

Single notes from the three sample collections and realistic monophonic phrases were mixed with white noise at different signal-to-noise ratios (SNRs). The results were very similar for both isolated notes and realistic phrases. With SNRs of 0 to 10 dB, almost all examples were classified as flutes; with a further decrease of noise to a SNR of 15 dB, a few violin tones were also correctly identified. This bias towards the flute model does not seem to be caused by a similarity between the noise and flute tones, as the noise alone resulted in such small probabilities from all models that they came within rounding error of 0. Recognition accuracy was only above chance at very high SNRs of 20dB, although with around 40% correctly identified examples the results were still well below those obtained for clean signals.

Making use of the missing feature approach based on *a priori* masks improved results significantly at all SNR levels. Averaging over all tested conditions, the use of missing feature masks improved recognition accuracy by 27% (Figure 3).

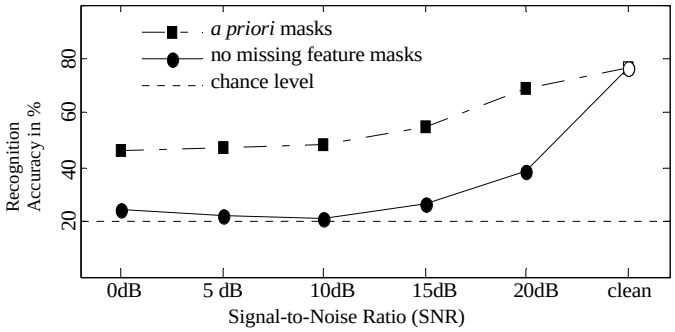

Figure 3: Recognition accuracy in the presence of white noise at various SNRs, with and without missing feature masks.

3.2.2 Two Concurrent Instrument Sounds

A priori masks were also used to identify the instruments in combinations of two independent monophonic examples, which were always played by different instruments. For each sample collection, a test set was derived by taking all possible combinations of two tones within one octave (C4-C5), excluding intervals in which both tones had the same F0 (3120 combinations per sample collection). Before mixing, the tones were normalised to have equal root-mean-square (rms) power. The length of each sound was determined by the shorter of the two tones to ensure that two instruments were present for the whole mixture. Average recognition accuracy was 59% for the McGill samples, 63% for the Ircam samples and 65% for the Iowa sample collection; representing an average drop in performance of less then 5% compared to the monophonic control condition. A confusion matrix averaging across the three examples is shown in Table 2.

stimulus \ response	Flute	Clarinet	Oboe	Violin	Cello
Flute	**73%**	3%	3%	12%	8%
Clarinet	22%	**47%**	10%	14%	7%
Oboe	1%	6%	**73%**	9%	11%
Violin	0%	3%	8%	**68%**	21%
Cello	3%	2%	15%	27%	**51%**

Table 2: Confusion matrix for mean instrument recognition of two concurrent notes using a priori masks.

The same approach was also tested with mixtures of realistic monophonic recordings. Using the same examples as for the monophonic condition, with all possible mixtures of two different instruments (500 different combinations), average recognition accuracy was 74%, a drop of 14% compared to the monophonic control condition. Generally the confusions for realistic phrases are very similar to those of isolated tone combinations. The main difference lies in the lower level of confusions between violin and cello for the realistic phrases. This is most likely due to the fact that for the isolated notes all examples were taken from the same octave, while the realistic phrases span the whole natural pitch range of the instruments.

An additional factor that could influence recognition accuracy is the interval relationship between the two notes. Critical intervals could be octaves and fifths, because the amount of overlap between partials from the target and the non-target tone is high. Recognition results for octaves were indeed about 10% below average, while no drop in performance occurred for fifths. Other intervals that could pose additional problems are seconds, where the F0s of the two notes are very close, and the individual partials might not be separated by the spectral features. Again, no drop in performance occurred, so the system proved to be quite robust towards the actual interval relationship of the notes.

3.3 Pitch-based Masks

As a next step towards realistic performance, we used combinations of two notes with masks based on the F0s of the non-target tone, which were in this case manually supplied to the algorithm. Since the system was shown earlier to be quite robust towards missing features, it seemed preferable to exclude too many features than to risk including corrupted features. Some preliminary tests supported this approach, as recognition using missing feature masks based on ‘broadened’ harmonics

improved recognition accuracy by up to 10%. The exact amount of deletions did not have a strong influence; a relative broadening of ±2.5% (slightly less than ± a quarter tone) worked well and was subsequently used for further experiments.

Recognition accuracy was 49% for the McGill samples, 43% for the Iowa and 48% for the Ircam samples. A confusion matrix averaging across the three conditions is shown in Table 3. All instruments were correctly identified in the majority of cases, except for the cello which was often mistaken for a violin. As all test tones were from the same octave and therefore relatively high for a cello, this confusion is not very surprising. Informal listening tests confirmed that low violin and high cello tones were hard to distinguish for humans, even in the clean monophonic condition were the system performed relatively well.

response / stimulus	Flute	Clarinet	Oboe	Violin	Cello
Flute	**54%**	4%	5%	27%	9%
Clarinet	25%	**44%**	7%	17%	7%
Oboe	17%	8%	**48%**	18%	8%
Violin	7%	3%	5%	**64%**	20%
Cello	16%	4%	15%	34%	**31%**

Table 3: Confusion matrix for mean instrument recognition of two concurrent notes with pitch based masks.

3.4 Bounded Marginalisation

For combinations of two instrument sounds, often more than half of the features are marked as unreliable and subsequently excluded from the recognition process. We now tested if the inclusion of the values of these unreliable features as upper bounds for the corresponding feature values could be used to improve recognition accuracy.

However, when tested with combinations of two isolated notes or monophonic phrases using *a priori* or pitch-based masks, no significant improvement was found. This result is at first rather unexpected, as bounded marginalisation can improve results significantly for speech recognition in the presence of noise (Cooke *et al.*, 2001). However, most noises used to test robust speech recognition systems are mainly inharmonic and therefore quite different from musical instrument tones. To see if the lack of improvement was due to this difference, we tested bounded marginalisation on monophonic sounds mixed with white noise at various SNRs. In these cases, the use of bounds did improve results considerably. For all mixtures with an SNR level between 0dB and 20dB, recognition accuracy using *a priori* masks and bounded marginalisation was on average as good as with clean monophonic signals.

The reason why the use of upper bounds proved only to be useful with random noise, but not with a harmonically structured tone, can probably be explained in terms of the different distribution of energy. With a musical tone, the energy is high at the frequencies where a harmonic overtone is present, and low otherwise. Frequency regions that are not excited by a partial of the interfering tone are therefore less likely to be marked as unreliable, while the energy in frequency regions where an interfering parital is present is likely to be well above the energy caused by the target tone alone.

Upper bounds appear to be useful only when the difference between observed energy (feature values) and energy caused by the target sound is relatively small, but in these cases bounded marginalisation improves the noise robustness of the recogniser considerably. Even though the white noise used in our experiments is an extreme case due to its complete spectral flatness, the use of bounded marginalisation could prove to be very useful for instrument recognition in noisy recordings. While pitch-based missing feature masks are not usable in these cases, various other noise estimation algorithms have been developed in the context of robust speech recogntion. They can be easily integrated with a missing feature approach and have been shown to lead to good results for speech mixed with various inharmonic noise sources (Cooke *et al.*, 2001).

4 Conclusions and Future Work

A system for the identification of musical instrument tones based on missing feature theory and a GMM classifier has been described. It generalises well, giving good results on single note recordings and on realistic musical phrases. Especially for the latter, results are well comparable to those of other systems directly designed for the identification of monophonic examples. Importantly, the system introduced here is not limited to identification of instruments in monophonic music. Rather, by using missing feature masks, the system is able to identify two different instruments playing concurrently. The use of missing feature masks also aids the recognition of monophonic instrument sounds in noisy conditions.

The system was primarily evaluated using *a priori* masks for combinations of isolated tones and independent monophonic phrases. Using pitch-based masks for isolated note combinations, performance was still good, but about 15% lower than with the use of *a priori* masks. This indicates that a more detailed analysis of which features are dominated by which source could lead to an improvement for the estimation of the pitch-based masks. Nevertheless, the system has been shown earlier (Eggink and Brown, 2003) to be able to reliable identify the instruments in a duet recording taken from a commercially available CD, using missing feature masks based on the F0s estimated by the system.

To be a useful tool for instrument classification tasks in the context of automatic transcription or musical information retrieval, not necessarily every note has to be correctly identified, as higher musical knowledge can help to suppress a few random errors. The integration of such higher level knowledge into our system will form part of our future work. Also, we intend to test how the system performs when more than two concurrent instruments are present. But the results achieved so far are encouraging, and it seems that our eventual goal - an automatic transcription system for audio recordings of classical chamber music played by small ensembles - is achievable.

Acknowledgments

JE is supported by the IHP HOARSE project. GJB is supported by EPSRC grant GR/R47400/01 and the MOSART IHP network.

References

Brown, J.C., Houix, O. & McAdams, S. (2001). Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the Acoustical Sociecty of America*, 109(3), 1064-1072

Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication* 34, 267-285

Drygajlo, A. & El-Maliki, M. (1998). Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-98*, 121-124

Eggink, J. & Brown, G.J. (2003). A missing feature approach to instrument identification in polyphonic music. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-03*, 553-556

Iowa Musical Instrument Samples, `http://theremin.music.uiowa.edu`

Ircam Studio Online (SOL), `http://www.ircam.fr`

Kashino, K. & Murase, H. (1999). A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication* 27, 337-349

Kinoshita, T., Sakai, S. & Tanaka, H. (1999). Musical sound source identification based on frequency component adaptation. *Proceedings IJCAI-99 Workshop on Computational Auditory Scene Analysis*, Stockholm, Sweden

Klapuri, A. (2001). Multipitch estimation and sound separation by the spectral smoothness principle. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-01*, 3381-3384

Marques, J. & Moreno, P. (1999). A study of musical instrument classification using Gaussian mixture models and support vector machines. *Cambridge Research Laboratory Technical Report Series CRL/4.*

Martin, K. (1999) *Sound-source recognition: A theory and computational model*. PhD Thesis, MIT.

Opolko, F. & Wapnick, J. (1987). *McGill University master samples* (CD), Montreal, Quebec: McGill University

Raphael, C. (2002). Automatic transciption of piano music. *Proceedings of the International Conference on Music Information Retrieval, ISMIR-02*